Journal of Digital Contents Society Vol. 17 No. 6 Dec. 2016(pp. 565-579)
https://doi.org/10.9728/dcs.2016.17.6.565

Cancer subtype's classifier based on Hybrid Samples Balanced Genetic Algorithm and Extreme Learning Machine

Vasily Sachnev*, Sundaram Suresh, Yong Soo Choi*****

Abstract

In this paper a novel cancer subtype's classifier based on Hybrid Samples Balanced Genetic Algorithm with Extreme Learning Machine (hSBGA-ELM) is presented. Proposed cancer subtype's classifier uses genes' expression data of 16063 genes from open Global Cancer Map (GCM) data base for accurate cancer subtype's classification. Proposed method efficiently classifies 14 subtypes of cancer (breast, prostate, lung, colorectal, lymphoma, bladder, melanoma, uterus, leukemia, renal, pancreas, ovary, mesothelioma and CNS). Proposed hSBGA-ELM unifies genes' selection procedure and cancer subtype's classification into one framework. Proposed Hybrid Samples Balanced Genetic Algorithm searches a reduced robust set of genes responsible for cancer subtype's classification from 16063 genes available in GCM data base. Selected reduced set of genes is used to build cancer subtype's classifier using Extreme Learning Machine (ELM). As a result, reduced set of robust genes guarantees stable generalization performance of the proposed cancer subtype's classifier. Proposed hSBGA-ELM discovers 95 genes probably responsible for cancer. Comparison with existing cancer subtype's classifiers clear indicates efficiency of the proposed method.

Keywords : Cancer Detection, Genetic Algorithm, Learning Machine, Classification

하이브리드 균형 표본 유전 알고리즘과 극한 기계학습에 기반한 암 아류형 분류기

Vasily Sachnev*, Sundaram Suresh, 최용수*****

요 약

본 논문에서는 극한 기계학습을 이용하는 하이브리드 균형 표본 유전자 알고리즘(hSBGA-ELM)을 기반으로 한 새로운 암 아류형 분류자를 제안하였다. 제안 된 암 아류형 분류자는 정확한 암 아류형 분류기 설계를 위해 공개 전체암지도 (Global Cancer Map)로부터 15063개의 유전자 발현 데이터를 사용합니다. 제안된 방법에서는 14가지(유방암, 전립선 암, 폐암, 대장 암, 림프종, 방광, 흑색 종, 자궁, 백혈병, 신장, 췌장, 난소, 중피종 및 CNS)의 암 아류형을 효율적으로 분류합니다. 제안 된 hSBGA-ELM은 유전자 선택 절차 및 암 아류형 분류를 하나의 프레임 워크로 단일화 한다. 제안 된 하이브리드 균형 표본 유전 알고리즘은 GCM 데이터베이스에서 이용 가능한 16,063 개의 유전자로부터 암 아류형 분류를 담당하는 축소된 강인 유전자 셋을 찾는다. 선택/축소된 유전자 세트는 익스트림 기계학습을 이용하여 암 아류형 분류기를 구성하는데 사용된다. 결과적으로, 크기가 축소된 강인 유전자 집합이 제안하는 암 아류형 분류기의 안정된 일반화 성능을 보장하게 한다. 제안 된 hSBGA-ELM은 암에 관여하는 것으로 예측되는 95개의 유전자를 발견하였으며 기존의 암 아류형 분류기와의 비교를 통해 제안 된 방법의 효율을 보여준다.

키워드 : 암 검출, 유전 알고리즘, 기계 학습, 분류

※ Corresponding Author : Yong Soo Choi
Received : December 12, 2016
Revised : December 27, 2016
Accepted : Decembe 31, 2016
* School of Information, Communication and Electronics Engineering, Catholic University
email: bassvasys@hotmail.com
** School of Computer Science and Engineering, Nanyang Technological University, Singapore
email: ssundaram@ntu.edu.sg

1. Introduction

Cancer is one of the most harmful and mostly untreatable diseases which is difficult to diagnose especially in early stages. Developed cancer diagnosis techniques are generally based on pathological tissues analysis, which is not always accurate and may result expensive and inefficient treatment. Hence, new diagnosis techniques with better accuracy based on new principals are needed.

New type of cancer diagnosis techniques based on analyzing genetic information has been proposed recently. Such genetic based cancer diagnosis techniques may efficiently diagnose cancer in early stages. Extensive research in the bio-informatics, micro-array techniques and machine learning significantly improves accuracy of genetic based cancer diagnosis techniques nowadays.

Micro-array techniques extract gene expression information from blade samples. Extracted genes'expression information is widely used in different research areas, including cancer subtype's diagnosis. Cancer diagnosis based on micro-array techniques utilizes genes' expression information to build an efficient classifier [5]. Selected genes are used as bio-markers for different type of cancer. However, there are several significant problems. a) Selecting an optimal subset of gene from an entire set is a big challenge, b) genes' expression data is usually highly unbalanced(few samples per class) and sparse (each sample has thousands of features). Such difficulties cause significant performance loss of the cancer subtype's classifiers based on genes' expression data.

Genetic based cancer diagnosis techniques exploit genes'expression information given in various data bases. The most famous is Global Cancer Map (GCM) [1]. GCM data base collects data from 6 medical institutes and contains genes' expression information of 16063 genes for 14 types of cancer. GCM genes' expression data base is widely used for experiments with genetic based cancer diagnosis techniques [2][3][4]. Koller et. al. [2] searched an optimal set of genes for accurate classification of 14 types of cancer. Lee et. al. [3] used genes' expression data from GCM data base to build classifier for ovarian cancer. Lin et. al. [4] created classifier for 14 types of cancer using GCM data base. Few other genes' expression data bases are exist in literature. Alizadeh et. al. [6] used genes' expression data base for diagnosing lymphoma. Pomeroy et. al. [7] collected genes' expression data base for embryonal central nervous tumor. West [8] investigated breast cancer.

Due to highly unbalanced gene's data set and high-dimensional feature space the choice of the proper machine learning technique is a big challenge.

Direct use of complete set of 16063 genes from GCM data based to train cancer subtype's classifier may not be efficient. Machine learning techniques presented in literature, in general, may not utilize high dimensional features efficiently. Researchers tried different variation of Support Vector Machine (SVM) [10], [11], [12], [13], traditional neural network [2], or fuzzy neural network [4]. Hence, new approach suitable to handle high dimensional features is needed.

Methods based on searching an optimal reduced setof genes resolve the problem. Reduced set of features is usually small enough for efficient training using machine learning techniques. Saraswathi et. al [13] used GCM data base and Integer Coded Genetic Algorithm with Particle Swarm Optimization to search a reduced set of genes

*** Division of Liberal Arts & Teaching, Sungkyul University
Tel: +82-31-467-8374
email: ciechoi@sungkyul.ac.kr

suitable for efficient classification 14 types of cancer. However, in the presented Integer Coded Genetic Algorithm the number of genes is assigned manually. Later Sachnev et. al. [14] proposed completely automatic cancer subtype's classifier based on Binary Coded Genetic Algorithm. Their method uses GCM data base and classifies14 types of cancer. Authors reported about 92 genes selected by BCGA, and 52 discovered bio-markers.

Proposed Hybrid Samples Balanced Genetic Algorithm coupled with Extreme Learning Machine (hSBGA-ELM) searches a reduced set of genes from GCM data base and builds an efficient cancer subtype's classifier for classifying 14 types of cancer. Proposed Hybrid Samples Balanced Genetic Algorithm (hSBGA) is a completely automatic approach for searching an optimal subset of genes from GCM data base. hSBGA is based on proposed genetic operators, such as Irregular Sample Balanced Crossover, Fixed Length Balanced Crossover and Sample Balanced Mutation. Proposed genetic operators designed specifically for GCM data and significantly simplified searching process. Generated by hSBGA reduced set of genes is used to build cancer subtype's classifier based on Extreme Learning Machine (ELM) [15]. Proposed hSBGA-ELM searches for a robust set of genes. Robust set of genes creates ELM classifier with stable generalization performances.

Proposed paper is organized follows: Section II presents GCM data base. In the Section III framework of the proposed hSBGA-ELM is presented in details. Experimental results are presented in Section IV. Section V concludes a paper.

2. Global Cancer Map (GCM)

Global Cancer Map (GCM) data set [1] has been collected from 6 medical institutes. GCM contains genes'expression information from 190 tumor's patients with 14 types of cancer (breast, prostate, lung, colorectal, lymphoma, bladder, melanoma, uterus, leukemia, renal, pancreas, ovary, mesothelioma and CNS). In GCM data base each patients/sample contains genes' expression information about 16063 genes. GCM data is heavily disbalanced (see Table 1). GCM data provides 144 samples for training and 46 samples for testing.

	Breast	Prostate	Lung	Colorectal	Lymphoma	Bladder	Melanoma	Uterus	Leukemia	Renal	Pancreas	Ovary	Mesothelioma	CNS
Training	8	8	8	8	16	8	8	8	24	8	8	8	8	16
Testing	3	2	3	3	6	3	2	2	6	3	3	3	3	4
Total	11	10	11	11	22	11	10	10	30	11	11	11	11	20

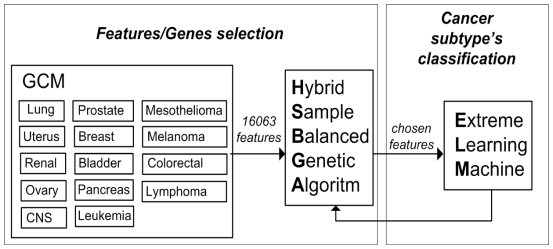
<Table 1> Distribution of cancer's types in GCM

Cancer subtype's classification problem classifies 190 patients/samples from GCM data base. Each patient/sample is presented as a vector of 16063 numbers. Cancer subtype classification problem contains 14 classes, where each class is a certain type of cancer. Distribution of samples per each classis presented in <Table 1>

3. Proposed hSBGA-ELM for cancer subtype's classification.

Proposed Hybrid Sample Balanced Genetic Algorithm with Extreme Learning Machine (hSBGA-ELM) creates efficient classifier for classifying 14 types of cancer using genes' expression data extracted from blade samples. Proposed hSBGA-ELM efficiently classifies following types of cancer: breast, prostate, lung, colorectal, lymphoma, bladder, melanoma, uterus, leukemia, renal, pancreas, ovary, mesothelioma and CNS. Framework of the proposed

hSBGA-ELM is displayed in Fig. 1. Proposed method consists of "Features/genes selection" and "cancer subtype's classification" steps (See Figure 1).



(Figure 1) Framework of the proposed hSBGA-ELM

Proposed hSBGA-ELM unifies "feature/genes selection" and "cancer subtype's classification" together intoone framework. hSBGA or "feature/genes selection" selects reduced set of features/genes from set of 16063 genes available in GCM data base. Selected genes are used to create an ELM classifier or "cancer subtype's classification"for cancer subtype's classification.

The detail explanations of the proposed hSBGA are presented below.

A. Hybrid Sample Balanced Genetic Algorithm(hSBGA)

Hybrid Sample Balanced Genetic Algorithm is a modification of famous Genetic Algorithm (GA) adapted for genes selection to solve cancer subtype's classification problem. Selected genes are used to create efficient Extreme Learning Machine classifier for classifying 14 types of cancer.

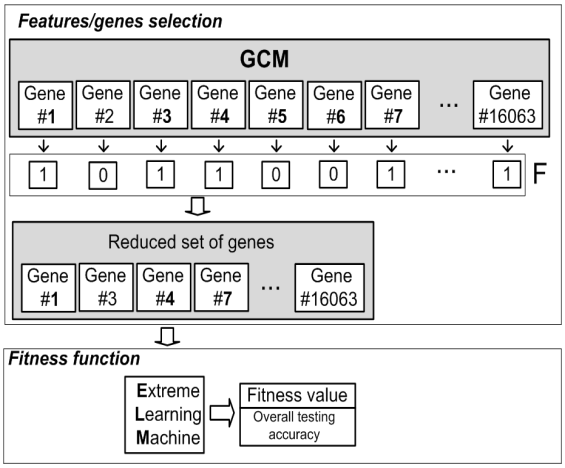
Genetic Algorithm (GA) is a famous tool for solving various optimization problems from engineering and science. GA is relatively simple and fast, GA is possible to adapt to various extremely complicated optimization problems with large number of variables and complex constraints. Sometimes only GA based methods may find suitable solution in an acceptable time. Thus, using GA for searching a reduced set of genes for cancer subtype's classification problem is a logical choice.

GA mimics few extremely important mechanisms of genomics from nature. In GA any optimization problem is transformed to a set of specifically assigned chromosomes. Each chromosome is a number from a predefined range, which has a significant impact to examined optimization problem. The set of meaningful chromosomes derives examined optimization problem in details and builds a solution for GA. Manipulations with chromosomes/numbers modify solutions, which affects examined optimization problem. Such modification may either improve or degrade a problem feedback. Manipulation with solutions which cause improvement helps GA to search for an optimal solution with maximum possible result. Each solution in GA is evaluated by fitness function. Fitness value is a numerical measure counted for each solution. Thus, GA numerically estimates solutions according to corresponding fitness values. Solutions with higher rank are used by GA again to generate other solutions which may have even better fitness values.

A string of binary coefficients is a set of meaningful chromosomes for proposed Hybrid Sample Balanced Genetic Algorithm.

Binary string is a set of binary "1" and "0" which manages a status of each gene from GCM data base for a given cancer subtype's classification problem. Thus, binary coefficient or each gene's status is a chromosome for a given problem. The set of 16063 binary coefficients builds a binary solution for a cancer subtype's classification problem. Binary coefficients "1" define a set of chosen genes for further analysis. Binary coefficients "0" define skipped genes. (See Figure 2). The set of chosen genes is used to build an ELM classifier for cancer subtypes' classification

problem.



(Figure 2) Features/genes selection and fitness function for cancer subtype's classification problem

Hybrid Samples Balanced Genetic Algorithm uses 3 proposed genetic operators: Irregular Sample Balanced Crossover, Fixed Length Balanced Crossover and Sample Balanced Mutation.

Genetic operators: Genetic Algorithm iteratively updates solutions by using genetic operators. Genetic operators modify solution's chromosomes similar to genes' exchange mechanism from nature. Genetic Algorithm is using 2 types of genetic operators: crossover and mutation. Crossover is responsible for genes' mixing in nature. Crossover from nature permutes genetic material from two sources (individuals) and creates a new genome for new source (individual). Finally, new individual with new genome contains genes' material from both sources. Genes' exchange mechanism is always chaotic and random. Due to crossover new individual contains properties from both sources (individuals). Such properties may be enhanced, degraded or kept same. Mutation in nature manipulates with genes randomly and creates insignificant genes' modifications. Such genes'modification may cause significant properties degradation, or sometimes improvements. Combination of crossover and mutation keeps a properties' balance in between input and new individuals.

Crossover and mutation in the Genetic Algorithm framework mimics genes' exchange mechanism from nature. GA crossover creates a new solution using chromosomes from 2 input solutions. In GA framework the genes' exchange strategy is always fixed. Note that, such genes' exchange strategy is problem specific and affects generalization performance of the GA. Thus, a choice of the proper crossover with suitable genes' exchange strategy is a big challenge. Proper choice of the mutation operation is also important.

Except a proper choice of the genetic operators (crossover and mutation) efficiency of the Genetic Algorithm depends on problem specification and Genetic Algorithm settings. Problem specification is responsible for transforming an examined optimization problem into a set of meaningful chromosomes. Incorrect transformation may lose some important problem's issues and may lead to significant performance degradation of the optimization process. Genetic Algorithm settings includes the number of solutions in each generation, crossover/mutation ratio and selection procedure settings. Incorrect settings destroy a balance between generations, damage a searching power of the GA and cause a significant performance lost.

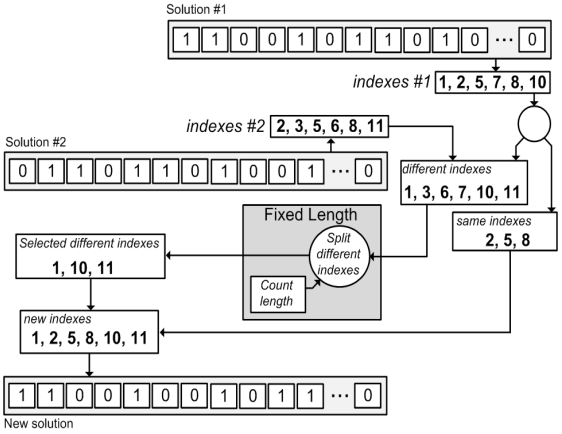
As was described before a proper choice of the genetic operators for certain optimization problem is always a big challenge. Various crossovers and mutation have been tested for searching an optimal reduced set of genes for cancer subtype's classification problem. All examined genetic operators do not enable an efficient search. Hence, new crossover suitable for an efficient search for cancer subtypes' classification problem has been

developed. Two new crossovers and one mutation developed specificallyfor cancer subtype's classification problem are presented in this paper: Fixed Length Balanced Crossover, Irregular Sample Balanced Crossover and Sample Balanced Mutation.

Well-known crossovers failed because of few reasons. Note that, binary coefficients link to a set of certain genes from GCM data base. Chosen set of genes is used to build ELM classifier which is then evaluated by overall testing efficiency. Thus, chosen set of genes play a vital role for GA. Wrong locations of binary "1" in a set of 16063 binary coefficients, may significantly damage performance of ELM classifier created using chosen genes and finally affect searching process of an entire GA. The number of binary "1"in each solution is relatively low, i.e., 20-200 coefficients in a set of 16063. All well-known crossovers create new solution with significant difference of the binary "1"between input and new solutions. Such difference significantly damage generalization performance of the ELM classifier and entire GA. Crossover may even generate solutions with all binary "0", which means no genes are chosen to build a training set for ELM classifier. Obviously it is unacceptable. Hence, two new crossovers: Fixed Length Balanced Crossover and Irregular Sample Balanced Crossover, and Sample Balanced Mutation are proposed in this paper.

Fixed Length Balanced Crossover is designed to hold the number of binary "1"in an acceptable level. Fixed Length Balanced Crossover creates a new solution by using only binary coefficients "1" taken from two input solutions. Fixed Length Balanced Crossover is displayed in (Figure 3).

Proposed Fixed Length Balanced Crossover collects locations of all binary "1" from "Solution #1" and "Solution #2" in sets "indexes #1" and "indexes #2" respectively. Proposed Fixed Length Balanced Crossover creates new solution with the fixed size of the set "new indexes" equal to average of sets'



(Figure 3) Fixed Length Balanced Crossover

sizes "indexes #1" and "indexes #2" (See Figure 3). Thus, Fixed Length Balanced Crossover never creates new solution with the number of binary "1"bigger or smaller compared to input "Solution #1" and "Solution #2. Locations of all binary "1" from "indexes #1" and "indexes #2" are then divided to "different indexes" and "same indexes" sets. Set "different indexes" keeps all different indexes; set "same indexes" keeps all common indexes. Set "new indexes" is created from randomly allocated indexes from set "different indexes" and all indexes listed in the set "same indexes". Lets $L_1, L_2, L_d, L_d^{'}, L_s$ and L_{new} are the size of sets "indexes #1" , "index #2", "different indexes", "selected different indexes", "same indexes" and "new indexes", respectively. Then $L_{new} = \lceil (L_1 + L_2) \cdot 0.5 \rceil$, $L_d^{'} = L_{new} - L_s$. Block "Split different indexes" picks $L_d^{'}$ indexes from set "different indexes" randomly and builds the set "Selected different indexes". Finally, set "new indexes" unifies indexes from sets "same indexes" and "Selected different indexes". Solution "New solution" is

created by placing binary "1" according to indexes in the set "new indexes".

An example of using proposed Fixed Length Balanced Crossover is presented in Fig. 3. Two input solutions "Solution #1" and "Solution #2" have binary coefficients "1" located in {1, 2, 5, 8, 10} (set "index #1") and {2, 3, 5, 6, 8, 11} (set "indexes #2"), respectively. Then, extracted indexes are divided into "same indexes" {2, 5, 8} and "different indexes" {1, 3, 6, 7, 10, 11}. The number of indexes in "Selected different indexes" is computed as follows: $L_d^{'} = L_{new} - L_s$, where $L_1 = 6, L_2 = 6, L_s = 6$ and

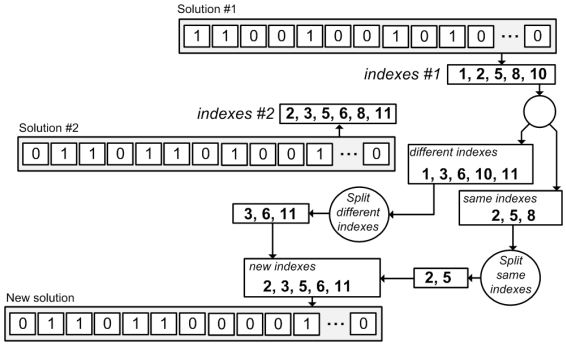
$L_{new} = \lceil (L_1 + L_2) \cdot \ 0.5 \rceil = \lceil (6+6) \cdot \ 0.5 \rceil = 6$

Then $L_d^{'} = L_{new} - L_s = 6 - 3 = 3$. Thus, block "Split different indexes" picks 3 indexes from the set "different indexes" randomly and builds a set "Selected different indexes" {1, 10, 11}. Set "new indexes" {1, 2, 5, 8, 10, 11} unifies set "same indexes" {2, 5, 8} and set "Selected different indexes" {1, 10, 11}. Finally, indexes from the set "new indexes" define positions of binary "1" for a "new solution". (see Figure 3)

Proposed Fixed Length Balanced Crossover controls the number of meaningful binary coefficients and may efficiently process any binary solutions for cancer subtype's classification problem. Fixed Length Balanced Crossover always creates new solution with number of binary "1" in an acceptable range.

Genetic Algorithm is working properly when the performances of the chosen genetic operators are balanced. Proposed Fixed Length Balanced Crossover never creates solutions with wrong number of binary "1", but at the same time it does not contain enough randomness for an efficient search. Hence, another crossover with less regulation is needed.

Irregular Samples Balanced Crossover is an improved version of Fixed Length Balanced Crossover designed to compensate enhanced regulation in Fixed Length Balanced Crossover. Irregular Samples Balanced Crossover creates new solution from randomly picked indexes from both "different indexes" and "same indexes" (see Fig 4.). Irregular Samples Balanced Crossover contains two random splitter "Split different indexes" and "Split same indexes". Random split parameter for splitter "Split different indexes" is $s_{dif} \in [0.2 \sim 0.8]$ random split parameter for splitter "Split same indexes" is $s_{same} \in [0.8 \sim 1.0]$. Every time when GA runs Irregular Samples Balanced Crossover parameters s_{diff} and s_{same} are updated.



(Figure 4) Irregular Samples Balanced Crossover

Proposed Irregular Samples Balanced Crossover balances GA searching procedure in many various scenarios as follows. In the situation then both input solutions have small number of common indexes, most of the binary "1" from both sources are located in the set "different indexes". New solution contains significant portion of the indexes from "different set" and the difference between new solution and input solutions is significant from the indexes point of view. Solutions mostly created from randomly chosen indexes may either degrade or improve properties, or keep its same. In the situation when both input solutions have many common binary "1", most of the indexes from both input solutions are

located in the set "same indexes". In this case new solution is mostly created from "same indexes". Thus, proposed Irregular Sample Balanced Crossover keeps convergence of the GA. When all solutions are mostly random, any new solution does not have many indexes in common and new solution is created from "different indexes". Such random manipulations with binary "1"is similar to chaotic search, which is very important in the beginning of the searching process when GA is searching promising solutions' areas with significant fitness values. In opposite situation, When new solution has many indexes from input solutions in common, the searching process is almost done and GA has to focus on searching a global optimum. New solution is always created from common "1" with insignificant modifications, which is similar to searching an optimal solution.

Example of the proposed Irregular Samples Balanced Crossover is displayed in (Figure 4). In this example random split parameters for splitters "Split different indexes" and "Split same indexes" are $s_{dif}=0.6$ and $s_{same}=0.8$, respectively. $L_{diff}=5$ and $L_{same}=3$. Crossover randomly picks $\lceil L_{dif} \cdot s_{dif} \rceil = \lceil 5 \cdot 0.6 \rceil = 3$ from "different indexes" and $\lceil L_{same} \cdot s_{same} \rceil = \lceil 3 \cdot 0.8 \rceil = 2$ from "same indexes". Set "new index" {2, 3, 5, 6, 11} combines 3 indexes from "different indexes" {3, 6, 11} and 2 indexes form "same indexes" {2, 5}.

Proposed Irregular Samples Balanced Crossover is more efficient when the searching process is almost done and significant portion of the generated solutions contain same indexes. Sometimes input solutions may contain the same indexes. In this situation Fixed Length Balanced Crossover creates new solution exactly the same as input solutions. Irregular Sample Balanced Crossover does not have this drawback and always generates new solution different from input solutions, even if both input solutions have the same indexes.

In this research the concept of Hybrid crossover is implemented. Hybrid crossover contains a pool of few crossovers suitable for examined optimization problem. The pool of crossovers may contain well-known crossovers, crossovers with novel design, or special crossovers. Hybrid crossover chooses one crossover randomly from the pool every time when GA needs it. Hybrid crossover creates new solutions using all crossovers listed in a pool. Hybrid crossover is very efficient when single crossover does not guarantee efficient search and combination of few crossovers is needed. In this research Hybrid crossover randomly selects either Fixed Length Sample Balanced Crossover or Irregular Sample Balanced Crossover.

Proposed Samples Balanced Mutation creates a new solution with randomly allocated binary coefficients "1". The number of binary "1" in new solution keeps the same with number of binary "1" in input solution.

Fitness function evaluates each solution generated by GA. Fitness function produces a special measure or fitness value. Fitness value numerically estimates and ranks each solution in GA. In GA framework solutions with significant fitness values are used to createslightly different solutions with may be even better fitness values. Solutions with insignificant fitness values are skipped.

Extreme Learning Machine classifier is a fitness function in the proposed hSBGA-ELM. Proposed Hybrid Sample Balanced Genetic Algorithm creates solution or string of binary coefficients which identify status of gene in GCM data base. Then, all genes with corresponding binary coefficient "1" are collected into a set of reduced genes for cancer subtype's classification problem. Reduced set of genes is used then to create 10 ELM classifiers using random parameters.

Due to strong unbalance between cancer types (see Table 1) in GCM data base, popular n-hold validation approach to balance ELM randomness is not efficient for given cancer subtypes classification problem. Hence, new validation approach is needed.

In this research we propose a 10-fold 10-split validation procedure for accurate estimation. In the proposed method each reduced set of genes is evaluated through extensive testing using various ELM classifiers. Suggested training/testing sets are unified together into one samples' data base (see Table 1 column "Total"). Unified samples are then divided into training/testing set 10 times randomly. Each training/testing set is evaluated through 10-fold validation. Thus, finally each reduced set of genes is used to create 100 ELM classifiers (10 ELM classifiers per each training/testing set). Average overall testing accuracy of all created 100 ELM classifiers is a fitness value for proposed hSBGA-ELM.

Selection procedure assigns special selection probability to each solution in GA according to it fitness value. Solutions with significant fitness value have higher chance to be selected to build a new solution by genetic operators.

An efficient geometric ranking method [16] is used as a selection procedure in the proposed hSBGA-ELM. Geometric ranking method sorts all solutions in descending order according to its fitness values and assigns selection probabilities as follows:

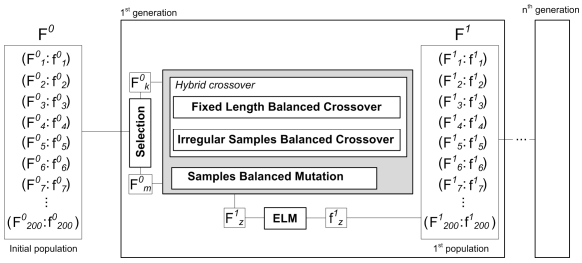
$$P_j = q'(1-q)^{r_j-1} \qquad (1)$$

where $q' = \dfrac{q}{1-(1-q)^N}$

q' is selection probability, r_j is a rank of j-th solution in the partially ordered set, and N is the population size. All details about geometric ranking method is given in [16]. In this research parameter $q = 10^{-3}$.

Termination criteria: Genetic Algorithm stops when GA no longer produces improvement during last 50 generations.

hSBGA-ELM framework: Proposed hSBGA-ELM starts form initialization step and processes several generations until termination criteria is satisfied (see Fig.5). Each generation proceeds selection procedure, genetic operators and fitness function. Initialization step starts from generating 200 binary solutions randomly. Each binary solution contains 16063 binary coefficients. The number of binary "1" is limited to 20-200. Then, each generated random solution from initial step is used to build 200 reduced sets of genes using GCM data base (see Figure 2). Each reduced set of genes is used to build 100 ELM classifier for proposed 10-fold 10-split validation procedure (see subsection "Fitness function"). Set of 100 created ELM classifiers is a fitness function and average overall testing accuracy is a fitness value. Finally, initialization step creates initial population F^0, which combines 200 generated binary solutions $\{F^0_1, F^0_2, F^0_3, ..., F^0_{200}\}$ and corresponding fitness values $\{f^0_1, f^0_2, f^0_3, ..., f^0_{200}\}$ (see Fig. 5). Any n-th generation creates 200 new solutions for n-th population F^n.



(Figure 5) Framework of the proposed hSBGA-ELM.

n-th generation starts from selection procedure based on geometric ranking method. For first 140 new solutions selection procedure

picks 2 input solutions from previous population F^n-1 according to assigned probabilities. Selected pair of binary solutions is processed by Hybrid crossover to create one new solution. Hybrid crossover randomly picks either Fixed Length Balanced Crossover or Irregular Sample Balanced Crossover. Rest 60 solutions in the population F^n are created by Sample Balanced Mutation. Finally, n-th population F^n is a combination of all binary solutions $\{F_1^0, F_2^0, F_3^0, ..., F_{200}^0\}$ and corresponding fitness values $\{f_1^0, f_2^0, f_3^0, ..., f_{200}^0\}$. In each generation crossover creates 70% or 140 new solutions, mutation creates rest 30% or 60 new solutions. GA creates generations one by one until termination criteria is satisfied.

B. Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) is machine learning technique with extremely fast learning, which is single hidden layer feed-forward neural network. For ELM controlled by Gaussian hidden neurons input weights and bias of the hidden neurons are randomly assigned and output weights are computed analytically [15, 17].

In the proposed hSBGA-ELM Extreme Learning Machine solves cancer subtype's classification problem. ELM classifier has been trained using 144 samples chosen randomly from the set of 190 samples available in GCM data base. Each sample contains n features/genes selected by hSBGA. Created ELM classifier is used then to classify 14 types of cancer.

ELM classifier is creates as follows: Given GCM data is divided to 144 samples for training and 46 samples for testing, i.e.
$\{(X_{tra}^1, c_{tra}^1), ..., (X_{tra}^t, c_{tra}^t)\}, (X_{tra}^{144}, c_{tra}^{144})$ and
$\{(X_{test}^1, c_{test}^1), ..., (X_{test}^t, c_{test}^t), ..., (X_{test}^{46}, c_{test}^{46})\}$.
where $X_{test}^t \in R^n$ and $X_{tra}^t \in R^n$ are n-dimensional feature vectors for t-th sample and $c^t = 1,2,...,14$ is class label. The coded class label $y_k^t \in R^{14}$is calculated as follows:

$$y_k^t = \begin{cases} 1, & \text{if } c^t = 1 \\ -1, otherwise \end{cases} \quad k = 1,2,...,14 \qquad (2)$$

Then $y_{tra} = \{y_{tra}^1, y_{tra}^2, ..., y_{tra}^{144}\}$ and $y_{test} = \{y_{test}^1, y_{test}^2, ..., y_{test}^{46}\}$ are sets of vectors with coded class labels for testing and training. $y_{tra} \in R^{14 \times 144}$, $y_{test} \in R^{14 \times 46}$.

Extreme Learning Machine is created as follows:

Training phase:

1) Assign number of hidden neurons L.

2) Generate sets of input weights $A \in R^{L \times n}$ and width (bias) $b \in R^{L \times 1}$ of hidden Gaussian neurons randomly.

3) Compute the hidden layer output matrix $G_{tra} \in R^{L \times 144}$

$$G_{tra} = \begin{pmatrix} g_1^1 \cdots g_1^{144} \\ \vdots \quad \vdots \\ g_L^1 \cdots g_L^{144} \end{pmatrix} \qquad (3)$$

where g_j^t is a response of j-th hidden neuron for t-th sample calculated as follows:

$$g_j^t = \exp\left(-\left(\frac{(X_{tra}^t - A_j)^T \cdot (X_{tra}^t - A_j)^T}{2 \cdot b_j^2}\right)\right) \qquad (4)$$

4) Compute output weights $\beta \in R^{L \times 1}$.

$$\beta = y_{tra} \cdot y_{tra}^+ \qquad (5)$$

where $+$ is a Moore-Penrose generalization inverse.

5) Compute predicted coded class labels

$$\hat{y}_{tra} = \beta \cdot G_{tra} \qquad (6)$$

6) Define predicted class label as follows:

$$\hat{c}_{tra}^k = \arg(\max(y_{tra})), k = 1,2,3,...,14 \qquad (7)$$

Testing phase:

1) Compute the hidden layer output matrix $G_{test} \in R^{L \times 46}$ as follows:

$$G_{tra} = \begin{pmatrix} g_1^1 \cdots g_1^{46} \\ \vdots \quad \vdots \\ g_L^1 \cdots g_L^{46} \end{pmatrix}$$

where

$$g_j^t = \exp\left(-\left(\frac{(X_{test}^t - A_j)^T \cdot (X_{test}^t - A_j)^T}{2 \cdot b_j^2}\right)\right) \quad (8)$$

2) Compute predicted coded class labels

$$\hat{y}_{test} = \beta \cdot G_{test} \quad (9)$$

3) Define predicted class label for testing as follows:

$$\hat{c}_{test}^k = \arg(\max(y_{test})), k = 1,2,3,...,14 \quad (10)$$

4. Experimental results

Cancer is one of Experiments with proposed hSBGA-ELM include: extensive search for a reduced set of genes probably responsible for cancer, training an efficient ELM classifier for classifying 14 types of cancer, experimental results analysis and comparison with existing techniques. Proposed Hybrid Sample Balanced Genetic Algorithm based on developed Fixed Length Crossover, Irregular Sample Balanced Crossover and Sample Balanced Mutation is designed specifically for searching a reduced set of genes probably responsible for cancer. Proposed cancer subtype's classifier is trained using Extreme Learning Machine, selected reduced set of genes and set of 144 randomly selected samples from GCM data base. Created ELM classifier is then tested using set of rest 46 samples from GCM data base. Generalization performances of the best created ELM classifier for classifying 14 types of cancer is presented and analyzed in this section.

Cancer subtype's classification problem has 14 classes or cancer's types. Each class has to be accurately classified by ELM classifier. Any misclassifications have to be analyzed and fixed. In this research a misclassification analysis is processed by using the concept of confusion matrixes. Confusion matrixes provide information about any misclassifications and correct classification in a matrix form. (See Tables 2 and 3). Position of each examined sample in confusion matrix is computed using actual and predicted class labels. <Table 2> presents "Training confusion matrix" <Table 3> presents "Testing confusion matrix". Both matrixes have 14X14 size. Vertical index represents actual class label; horizontal index represents predicted class label. "Training confusion matrix" (see Table 2) has only 3 misclassified samples: one lymphoma sample has been classified as ovary, one leukemia sample has been classified as colorectal and one CNS sample has been classified as lung. "Testing confusion matrix" (see Table 3) has only 2 misclassified samples: one lymphoma sample has been classified as lung and one CNS sample has been classified as bladder. Overall testing accuracy is $\eta^{test} = 95.45\%$. Overall training accuracy is $\eta^{train} = 97.92\%$.

Average overall testing accuracy computed using 100 ELM classifiers for 10-fold 10-split validation is We also discovered a set of 95 best genes, which is used to create ELM

Actual class labels

Coded class labels	Breast	Prostate	Lung	Colorectal	Lymphoma	Bladder	Melanoma	Uterus	Leukemia	Renal	Pancreas	Ovary	Mesothelioma	CNS
Breast	8	0	0	0	0	0	0	0	0	0	0	0	0	0
Prostate	0	8	0	0	0	0	0	0	0	0	0	0	0	0
Lung	0	0	8	0	0	0	0	0	0	0	0	0	0	0
Colorectal	0	0	0	8	0	0	0	0	0	0	0	0	0	0
Lymphoma	0	0	0	0	15	0	0	0	0	0	0	**1**	0	0
Bladder	0	0	0	0	0	8	0	0	0	0	0	0	0	0
Melanoma	0	0	0	0	0	0	8	0	0	0	0	0	0	0
Uterus	0	0	0	0	0	0	0	8	0	0	0	0	0	0
Leukemia	0	0	0	**1**	0	0	0	0	23	0	0	0	0	0
Renal	0	0	0	0	0	0	0	0	0	8	0	0	0	0
Pancreas	0	0	0	0	0	0	0	0	0	0	8	0	0	0
Ovary	0	0	0	0	0	0	0	0	0	0	0	8	0	0
Mesothelioma	0	0	0	0	0	0	0	0	0	0	0	0	8	0
CNS	0	0	**1**	0	0	0	0	0	0	0	0	0	0	15

<Table 2> Training confusion matrix

classifier with the best overall training and testing accuracies and has the best average overall testing accuracy for 10-fold 10-split validation. $\hat{\eta}^{test} = 91.2\%$.

Coded class labels \ Actual class labels	Breast	Prostate	Lung	Colorectal	Lymphoma	Bladder	Melanoma	Uterus	Leukemia	Renal	Pancreas	Ovary	Mesothelioma	CNS
Breast	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Prostate	0	2	0	0	0	0	0	0	0	0	0	0	0	0
Lung	0	0	3	0	0	0	0	0	0	0	0	0	0	0
Colorectal	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Lymphoma	0	0	1	0	5	0	0	0	0	0	0	0	0	0
Bladder	0	0	0	0	0	3	0	0	0	0	0	0	0	0
Melanoma	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Uterus	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Leukemia	0	0	0	0	0	0	0	0	6	0	0	0	0	0
Renal	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Pancreas	0	0	0	0	0	0	0	0	0	0	3	0	0	0
Ovary	0	0	0	0	0	0	0	0	0	0	0	3	0	0
Mesothelioma	0	0	0	0	0	0	0	0	0	0	0	0	3	0
CNS	0	0	0	0	0	1	0	0	0	0	0	0	0	3

<Table 3> confusion matrix.

We also discovered a set of 95 best genes, which is used to create ELM classifier with the best overall training and testing accuracies and has the best average overall testing accuracy for 10-fold 10-split validation.

4.1. Comparison with existing methods

Examined existing techniques should use GCM genes' expression data base and classify 14 types of cancer. Two existing techniques were selected: ICGA-PCO-ELM technique presented by Saraswathi et. al [13], and BCGA-ELM technique presented by Sachnev et. al. [14]. Both examined techniques and presented approach utilize same strategy: search for a reduced set of genes using different variation of Genetic Algorithm and training ELM classifier based on selected genes for accurate classifying cancer's types. ICGA-PCO-ELM [13] is semiautomatic and needs to select a number of genes manually. BCGA-ELM [14] is completely automatic. Maximum possible classification accuracy for two examined methods and proposed hSBGA-ELM is closed to 100%. The number of samples in testing set is just 46 and each misclassified sample significantly decreases classification accuracy. Comparison over average testing accuracies is more reliable (see Table 4).

	hSBGA-ELM	**ICGA-PCO-ELM**	**BCGA-ELM**
Training accuracy	**97.92**	96.0	-
Testing accuracy	**95.45**	98	95.45
Average testing accuracy	**91.2**	91	-

<Table 4> Classification accuracies

Proposed hSBGA-ELM shows 0.2% better average testing accuracy compared to ICGA-PCO-ELM. ICGA-PCO-ELM used 10-fold validation, which means that average testing accuracy has been computed using 10 testing accuracies from 10 ELM classifiers, whereas proposed hSBGA-ELM uses 10-fold 10 split validation, which collects testing accuracies form 100 ELM classifiers. Note that average testing accuracy is more important compared to maximum testing efficiency and may determine bio-markers responsible for cancer. Higher testing accuracy (98%) for ICGA-PCO-ELM means that there is only 1 misclassified sample among 46 available for testing (proposed method has 2 misclassified samples). Training accuracy of the proposed method is 1.92% higher compared to ICGA-PCO-ELM. Compared to BCGA-ELM presented in [14], proposed method shows similar overall testing accuracy ≈ 95%.

Authors in [14] did not report about average testing accuracy and training accuracy. Thus, direct comparison using those parameters is not possible.

5. Conclusion

In this paper an efficient cancer subtype's classifier based on proposed Hybrid Sample Balanced Genetic Algorithm with Extreme Learning Machine is presented. Proposed Hybrid Sample Balanced Genetic Algorithm uses proposed Fixed Length Crossover and Irregular Sample Balanced Crossover for an efficient search for reduced set of genes with promising properties. Reduced set of genes is used to build cancer subtype's classifier using Extreme Learning Machine. Experiments with existing techniques clear indicate efficiency of the proposed method.

Finally discovered reduced set of 95 genes, which was used to build an efficient cancer subtype's classifier, may help to better understand the basic hidden mechanisms of cancer. Such information may help to design new guidelines or drugs for cancer's treatment.

Acknowledgement

This work was supported by Catholic University of Korea, Research Funds 2016

References

[1] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures", Proceeding of National Academic Science US, vol. 98, no. 26, pp. 15149-15154, 2001. [

[2] D. Koller and M. Sahami, "Toward optimal feature selection," In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 284 - 292, Bari, Italy, 1996.

[3] Z. J Lee, "An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer," Artificial Intelligent in Medicine, vol. 42, no. 1, pp. 81-93, 2008.

[4] T.-C. Lin, R.-S. Liu, Y.-T. Chao, and S.-Y. Chen, "Multiclass Microarray Data Classification Using GA/ANN Method," in PRICAI 2006: Trends in Artificial Intelligence, vol. 4099, pp. 1037-1041, 2006.

[5] G. Piatetsky-Shapiro, P. Tamayo, K. Dnuggets, and U.M. Lowell, "Microarray Data Mining: Facing the Challenges," SIGKDD Explorations, vol. 5, no. 2, pp. 1-5, Dec. 2003.

[6] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol. 403, no. 6769, pp. 503- 511, 2000.

[7] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub, "Prediction of central nervous system embryonal tumor outcome based on gene expression," Nature, vol. 415, no. 6870, pp. 436-442, 2002.

[8] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, Jr., J.R. Marks, and J.R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," Proceedingof National Academic Science USA, vol. 98, no. 20, pp. 11462-11467, 2001.

[9] S. Saraswathi, S. Sundaram, N. Sundararajan, M.

Zimmermann, and M. Nilsen-Hamilton, "ICGA-PSO-ELM approach for accurate multiclass cancer classification resulting in reduced gene sets in which genes encoding secreted proteins are highly represented", IEEE ACM Transaction on Computational Biology and Bioinformatics, vol. 8, No. 3, pp. 452 - 463, 2011.

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," Machine Learning, vol. 46, no. 1-3, pp. 389-422, 2002.

[11] X. Zhou and D. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data," Bioinformatics, vol. 23, no. 9, pp. 1106-1114, 2007.

[12] 5. Y. Tang, Y.-Q. Zhang, and Z. Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," IEEE/ACM Transaction Computational Biology and Bioinformatics, vol. 4, no. 3, pp. 365-381, 2007.

[13] 6. Y. Wang, I.V. Tetko, H.A. Mark, E. Frank, A. Facius, K.F.X. Mayer, and H.W. Mewes, "Gene selection from microarray data for cancer classification - a machine learning approach," Computational Biology and Chemistry, vol. 29, no. 1, p. 37-46, Feb. 2005.

[14] Vasily Sachnev, Saras Saraswathi, Rashid Niaz, Andrzej Kloczkowski and Sundaram Suresh, "Multi-class BCGA-ELM based classifier that identifies biomarkers associated with hallmarks of cancer", BMC Bioinformatics, vol. 16, no. 166, 2015

[15] G.-B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications", Neurocomputing, vol. 70, no. 1-3, pp. 985990, 2006.

[16] S. Suresh, S. N. Omkar, V. Mani, T. N. G. Prakash, "Lift coefficient prediction at high angle of attack using recurrent neural network", Aerospace Science and Technology, vol. 7, pp. 595 602, 2003

[17] L. V. Ma, S. H. Park, J. H. Jang and J. H. Park, "Fuzzy Decision Making-based Recommendation Channel System using the Social Network Database, " J. of Digital Contents Society, Vol.17, No.5, 2016

Vasily Sacnev

2002년 : Komsomolsk-na-Amure State Tech. University, (B.S)
2004년 : Komsomolsk-na-Amure State Tech. University, (M.S)
2009년 : Korea University (PhD)
2010년～현재: Catholic University, Assistant Professor
관심분야 : Multimedia Security, Steganography, Steganalysis, Machine learning and Bio-informatics

Sundaram Suresh

1999년 : Bharathiyar University, INDIA(B.E)
2001년 : Indian Inst. of Science Bangalore, INDIA(M.E)
2005년 : Indian Inst. of Science Bangalore, INDIA (Ph.D)
2005년～2007: Nanyang Tech. University, post-doctoral
2007년～2008: Indian Inst. of Tech. -Delhi, Assistant Professor
2010년～현재: Nanyang Tech. University, Assistant Professor
관심분야 : Computational cognitive system, Neural networks, Intelligent control, Medical image processing, Mathematical optimization and Game theory

최 용 수

1998년 강원대학교
제어계측공학과 공학사
2000년 강원대학교
제어계측공학과 공학석사
2006년 강원대학교
제어계측공학과 공학박사
2006년～2007년 연세대학교 첨단융합건설연구단 연구교수
2007년～2013년 고려대학교 정보보호대학원 연구교수
2013년～현재 성결대학교 교양교직부 (멀티미디어) 조교수
관심분야 : Multimedia Hashing, Information Hiding, Watermarking, Steganography, Image Forensics, Forgery Detection 등